Enquête de cohorte et analyse multivariée : une analyse épistémologique et historique du rôle fondateur de l'étude de Framingham

Elodie Giroux

▶ **To cite this version:**

Elodie Giroux. Enquête de cohorte et analyse multivariée : une analyse épistémologique et historique du rôle fondateur de l'étude de Framingham. Epidemiology and Public Health = Revue d'Epidémiologie et de Santé Publique, 2008, 56 (3), pp.177-188. 10.1016/j.respe.2008.02.110 . halshs-00791124

HAL Id: halshs-00791124
https://shs.hal.science/halshs-00791124

Submitted on 6 Sep 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Enquête de cohorte et analyse multivariée : une analyse épistémologique et historique du rôle fondateur de l'étude de Framingham

Cohort study and multivariate analysis : an epistemological and historical analysis of the Framingham Heart Study

E. GIROUX
Université de Lyon 1
Institut d'Histoire et de Philosophie des Sciences et des Techniques (Paris)
Office of NIH History, Bethesda, Md, USA
elodie.giroux@univ-lyon3.fr

Article publié dans Revue d'Epidémiologie et de santé publique 56 (2008) 177–188
Site de la revue :
http://www.elsevier-masson.fr/product.jsp?id=EHS_FR_BS-PK-710&ptr=3376&gclid=CK7M_c3QtrkCFcaWtAod1VYANA

Begun in 1947 and still ongoing, the epidemiological study of heart disease known as the Framingham Study was one of the first prospective studies based on a large cohort and has rapidly been considered as the prototype and model of the cohort study. Nevertheless, an examination of its history reveals that the protocol does not at all correspond to today's standards for this type of study. How, then, can we account for the remarkable reputation of this study?

This paper consists in an epistemological and historical analysis of the Framingham Study that provides some of the answers to this question. In my treatment of the study's methodology, I focus on the issue of how the study population was constituted, and the manner in which the multiple factor analyses were conducted two issues that are now central to cohort studies and more generally to analytic epidemiology.

Thus, I show how the study population of Framingham and its long-term follow up have contributed significantly to the interpretation of the cohort as a sort of 'population-laboratory'. The data generated by this study, which have been very widely used by epidemiologists and other researchers, are unparalleled in terms of the amount of detailed clinical information available for such a long follow-up period. Furthermore, multivariate statistical modelling, which has become a standard statistical tool for clinical as well as epidemiological studies was introduced in the context of this study to improve the identification of significant factors in the simultaneous analysis of multiple correlations. Multivariate analysis has since proved crucial in shaping the epidemiological concept of 'risk factor' and in analysing multifactorial disease. Indeed, I suggest that the modern idea of multifactorial disease depends on the adaptation of this statistical method.

Thus, the Framingham Study played a leading role not only in remodelling epidemiology after the Second World War, in particular because of its contribution to the establishment of the cohort study as a standard method of investigation in etiological research, but also in constituting the 'risk factor approach' to disease.

History of medicine. Cohort studies. Prospective studies. Research design. Cardiovascular diseases. Multivariate Analysis. Risk Factors.

Commencée en 1947, l'enquête épidémiologique dite « de Framingham » sur les maladies coronariennes dure encore aujourd'hui. Elle est habituellement considérée comme l'une des premières grandes enquêtes prospective de cohorte et comme constituant son prototype.

Pourtant, si l'on se penche sur son histoire, son protocole d'étude est bien loin de correspondre aux exigences attribuées aujourd'hui à ce type d'enquête. Comment dès lors expliquer cette renommée ?

Cet article de nature épistémologique et historique propose d'apporter quelques éléments de réponse à cette question, principalement à partir de l'analyse de la méthode de l'enquête. Nous nous limitons à deux aspects devenus aujourd'hui fondamentaux pour ce protocole d'étude : la constitution de la cohorte et l'analyse de la corrélation multiple.

Nous montrons que, d'une part, la population d'étude et son suivi ont conduit à faire de la cohorte de Framingham une sorte de 'population laboratoire'. Ces données, très sollicitées pour diverses analyses, n'ont pas d'équivalent par leur durée et leur précision clinique. D'autre part, c'est à l'occasion des analyses de corrélation réalisées pour cette étude que les modèles statistiques multivariés furent introduits et adaptés à l'épidémiologie. Ces modèles ont été déterminants pour la constitution de la notion épidémiologique de 'facteur de risque' et l'analyse de l'origine multifactorielle des maladies. En effet, il nous semble que la conception moderne de la multifactorialité des maladies est tributaire de l'usage de cette méthode statistique.

Ainsi, au milieu du 20^{e} siècle, l'étude de Framingham est bien une enquête fondatrice à cause de sa contribution à l'émergence du rôle nouveau de l'enquête de cohorte dans la recherche étiologique et à la constitution d'une 'approche facteurs de risque' de la maladie.

Histoire de la médecine. Enquêtes de cohorte. Enquêtes prospectives. Plan d'étude. Maladies cardiovasculaires. Analyse multivariée. Facteurs de risque.

INTRODUCTION

C'est dans un contexte de transition épidémiologique des pays développés où les maladies dites chroniques, principalement les cancers et les maladies coronariennes, prenaient le pas sur les maladies infectieuses que l'épidémiologie développa deux principaux types d'étude de population ayant pour visée l'analyse étiologique. L'étude cas-témoins et l'étude de cohorte se sont en effet plus particulièrement constituées comme entité méthodologique au lendemain de la Seconde Guerre mondiale, aux États-Unis et en Grande-Bretagne [1]. C'est un tournant dans l'histoire de l'épidémiologie : la recherche étiologique des maladies y acquiert un rôle et une importance nouvelle aux côtés de la recherche expérimentale et clinique. Il importe d'élucider, à travers l'histoire de la constitution de ces enquêtes, comment ce rôle de l'épidémiologie dans l'analyse étiologique s'est développé.

L'étude prospective de cohorte, quand elle est possible, est aujourd'hui considérée comme le meilleur plan d'étude pour l'analyse étiologique et l'identification de facteurs de risque de maladies. Sa force relativement à l'étude cas-témoins est de permettre une mesure du taux d'incidence grâce à sa dimension prospective et ainsi, un calcul des risques relatifs, mais aussi de respecter la temporalité de la relation entre le facteur et son effet et de limiter les biais. Sa faiblesse est d'être très coûteuse à cause de sa longue durée et du grand nombre d'individus requis pour obtenir des associations significatives. L'origine de ce plan d'étude qui, à la différence de celle de l'enquête cas-témoins, n'est pas si aisément décrite comme un prolongement des études cliniques de séries de cas [1, 2], est souvent associée à l'enquête de Framingham sur les maladies coronariennes. Commencée en 1947 dans une ville près de Boston, cette étude fut très vite décrite par les épidémiologistes comme l'une des premières grandes enquêtes prospectives de cohorte et citée comme son modèle paradigmatique [3-6].

Or, ce statut d'étude pionnière et exemplaire suscite l'interrogation. Premièrement, des travaux récents en histoire de l'épidémiologie ont mis en évidence qu'il y eut des plans d'étude du type de l'enquête de cohorte bien avant celle de Framingham, tout au moins sous une modalité rétrospective [7-9] : notamment sur la tuberculose, en Allemagne avec l'étude de W. Weinberg en 1913 [10] et aux États-Unis avec les travaux de Wade H. Frost [11]. Par ailleurs, d'autres études prospectives de cohorte commencèrent de manière quasiment simultanée à celle de Framingham dans divers États des États-Unis. Deuxièmement, d'un point de vue méthodologique, comme l'a fait remarquer l'épidémiologiste américain Mervyn Susser, il y a une sorte de paradoxe dans la renommée de l'étude de Framingham : si l'on considère attentivement sa planification initiale, aujourd'hui les *peer reviewers* ne considèreraient pas qu'elle respecte le protocole de l'enquête de cohorte dont elle serait

pourtant l'origine [6]. Sa mise en place a été pour le moins tâtonnante, résultat de constructions et adaptations successives [12, 13]. Comment dès lors expliquer sa renommée et son statut de paradigme ?

L'objectif de cet article est d'apporter des éléments de réponse à partir d'une analyse historique et épistémologique de l'enquête. Dans l'espace imparti, nous ne saurions développer une analyse exhaustive de son histoire dans ses dimensions sociales, politiques et institutionnelles [13-18]. Nous nous limitons à deux éléments de nature méthodologique, devenus aujourd'hui essentiels à ce type d'étude : la constitution de la cohorte et l'analyse de la corrélation multiple. Nous pensons qu'ils justifient en grande partie le succès de l'enquête de Framingham. À travers l'analyse de cette enquête particulière, nous n'avons pas la prétention de faire l'histoire de l'étude de cohorte ou de 'l'approche facteurs de risque des maladie' mais en montrant le rôle qu'elle y a joué, nous mettons en lumière quelques éléments importants de cette histoire.

Avant de concentrer notre attention sur cette enquête, trois points doivent être rapidement évoqués. Premièrement, au milieu du 20^{e} siècle, les maladies coronariennes étaient devenues un véritable problème de santé publique aux États-Unis. Dès les années 1920, des cliniciens, des professionnels de santé publique et surtout, les assureurs avaient pris conscience du poids nouveau de ces maladies dans la mortalité générale [16, 19]. D'après les statistiques publiques de mortalité en 1950, ces maladies représentaient la première cause de mortalité : plus de 40% des décès leur étaient imputés [20]. Elles frappaient les esprits par la brutalité de leur manifestation et par le fait qu'elles touchaient principalement les hommes dans la pleine force de l'âge [21]. Au lendemain de la Seconde Guerre mondiale, l'*United States Public Health Service* (*US PHS*) déplaça son principal champ d'investigation des maladies infectieuses vers les maladies chroniques. Le besoin se faisait ressentir de disposer de meilleures statistiques de morbidité sur ces maladies ; les statistiques de mortalité étaient insuffisantes pour obtenir une évaluation rigoureuse de leurs taux de prévalence et d'incidence.

Deuxièmement, les maladies coronariennes faisaient de plus en plus l'objet de recherches cliniques et expérimentales sans que cela n'aboutissent toutefois à de réelles perspectives d'élucidation de leur pathogenèse. Grâce aux progrès réalisés au début du 20^{e} siècle dans la connaissance et la classification nosologique de ces maladies, la mise en œuvre de programmes de dépistage ou d'études de morbidité dans des populations délimitées devenait envisageable. Les travaux du médecin américain James Herrick (1861-1954) révélèrent que derrière l'apparence foudroyante et brutale de l'infarctus du myocarde se

dissimulait une maladie au développement progressif et silencieux ; Herrick livra une description clinique de ce qui devint une entité nosologique [22]. Cette description ajoutée à l'utilisation de l'électrocardiogramme facilitait le diagnostic de cette maladie du vivant du malade et la recherche clinique [23]. Par ailleurs, des découvertes en pathologie expérimentale conduisirent à mettre en évidence le rôle du dépôt de cholestérol dans le développement de l'athérosclérose [24]. Le mode d'alimentation américain commençait à être sérieusement mis en cause, suite notamment à la comparaison des statistiques de mortalité de ces maladies entre différents pays [25]. Les liens entre le mode d'alimentation, le taux de cholestérol et le développement d'athérosclérose faisaient l'objet de vives controverses et la recherche expérimentale laissait de nombreuses questions irrésolues. Toutes ces avancées contribuaient à écarter progressivement l'idée selon laquelle ces maladie étaient dues au vieillissement (« maladie dégénérative ») et apportaient un certain nombre d'hypothèses étiologiques dont les cliniciens et épidémiologistes allaient pouvoir s'emparer dans le cadre d'une enquête épidémiologique.

Un troisième point concerne le lien particulier de l'enquête prospective de cohorte avec les maladies coronariennes : ce plan d'étude fut d'emblée privilégié sur l'étude cas-témoins à la différence de ce qui se passait alors au même moment pour l'épidémiologie des cancers. La nature silencieuse et progressive de ces maladies, les difficultés rencontrées pour leur définition et leur diagnostic, plus importantes encore que pour les cancers en dépit des avancées précédemment évoquées, une conscience déjà grande de la part des cliniciens d'une étiologie complexe, mais aussi leur plus grand taux d'incidence, peuvent contribuer à expliquer cette particularité [13, 14].

LA POPULATION D'ETUDE : « UNE POPULATION LABORATOIRE »

La difficulté de l'enquête étiologique en épidémiologie est de travailler sur des données recueillies dans un contexte non expérimental. L'essentiel est alors de contrôler et encadrer le plus possible l'observation pour limiter les risques de biais (sélection, classement et confusion). La rigueur avec laquelle la population d'étude, le 'laboratoire' de l'épidémiologiste, est constituée est le principal moyen de contrôler et limiter les biais de sélection. C'est d'elle que dépend ensuite la qualité de l'analyse des données. Alors que le plan d'étude de cohorte et la notion de biais de sélection n'étaient pas encore entièrement formalisés au moment de la naissance de l'enquête de Framingham, nous examinons tout

d'abord, comment la définition de la population (qu'on appelle aujourd'hui la « population source ») fut abordée par les investigateurs et ensuite, la façon dont la cohorte fut constituée.

LA POPULATION SOURCE

La population générale d'une ville

Le choix, pour cette enquête, d'une population dite « générale et non sélectionnée » en retenant comme population source « l'ensemble de la communauté » [4] d'une ville, hommes et femmes, est l'un de ses traits les plus originaux. Notons toutefois d'emblée que la population d'une ville est toujours 'sélectionnée' d'une certaine manière, du simple fait qu'on choisit une ville particulière. Par ailleurs, on retient conventionnellement une tranche d'âge donnée comme source de l'échantillonnage et non pas la population entière de la ville. L'inclusion des femmes – plus de la moitié de la cohorte finale (2873 sur un total de 5209 individus) – est particulièrement étonnante dans la mesure où la maladie coronarienne touche principalement les hommes [21]. Pour un tel choix, l'enquête de Framingham fait quasiment figure d'exception. Hormis l'étude de Tecumseh (Michigan) qui commença dix ans plus tard (1959) dans une communauté mais avec un objectif explicitement plus 'écologique' [26, 27] d'étude d'une population dans son environnement physique, biologique et social et sur un ensemble plus large de maladies chroniques [28, 29], les études de cohorte sur les maladies cardiovasculaires qui débutèrent en même temps à Minneapolis (A. Keys) [30] puis, un peu plus tard, à Albany (J.T. Doyle) [31], à Los Angeles (J. Chapman) [32], à Chicago (les études de J. Stamler [33] et de O. Paul [34]) en Grande-Bretagne (J.N. Morris) [35, 36] sont menées à partir de sous-groupes professionnels. Seule l'étude de Chapman intègre des femmes mais en nombre très inférieur aux individus de sexe masculin. L'avantage, d'une part, de sous-groupes professionnels ou d'assurés est qu'il est beaucoup plus facile de recueillir les données et de garantir un bon suivi de la cohorte. Mais cette sélection peut introduire des biais importants. D'autre part, se limiter à la population masculine dans laquelle la maladie est la plus fréquente permet d'obtenir plus rapidement des résultats puisque l'analyse comparative dans une étude prospective de cohorte dépend de la fréquence des nouveaux cas. Comment dès lors expliquer le choix d'une « population générale » et d'une « approche communautaire » [37] à Framingham ? La notion de 'communauté' est ici troublante dans un contexte américain ; elle évoque l'idée d'un groupe ethnique ou d'une approche sociologique ou, tout au moins, 'écologique' au sens où nous venons de l'évoquer à propos de l'étude de Tecumseh.

Les débuts de l'enquête et en particulier, ses premiers objectifs, peuvent contribuer à fournir quelques explications. L'enquête naît en 1947 dans le cadre de l'*US PHS* avant d'être transférée en 1949 au tout nouveau *National Heart Institute* (*NHI*). Cette appartenance à l'État fédéral la distingue d'ailleurs des autres études de cohorte sur ces maladies qui furent menées à partir d'initiatives et d'institutions locales. En 1947, Gilcin Meadors, le premier responsable et concepteur de l'enquête, parle d'une « population 'normale' et non sélectionnée » par opposition aux « populations artificielles » (Meadors, 19 juillet 1947, Justification for Budget Estimate for the Sub-Project 'Epidemiology', *Memorandum,* Framingham Archives). Il souligne que dans les sous-groupes professionnels ou d'assurés, il est difficile d'obtenir des données valides pour le calcul des taux de prévalence et d'incidence des maladies coronariennes. De telles données sont présentées en 1948 comme un préalable nécessaire à toute comparaison éventuelle : « Il n'y a pas de données disponibles sur une population américaine moyenne ou typique avec laquelle des comparaisons pourraient être faites pour déterminer si les taux diffèrent (…). » (Meadors, Proposed Study of the Epidemiology or Cardiovascular Diseases, Rough Draft, 1948, Framingham Archives). On s'aperçoit ici que quand l'étude débute en 1947, l'objectif est premièrement celui de l'obtention de meilleures données sur la prévalence et l'incidence des maladies coronariennes dans la population américaine, dans un contexte de santé publique. La visée de recherche étiologique est présente mais de manière secondaire – comme un corrélat de l'étude d'incidence – et encore peu définie. Ce n'est que dans l'étape de définition du projet de suivi pour la mesure du taux d'incidence en 1949, quand l'étude de prévalence a bien avancé, que l'étude de Framingham devient explicitement et principalement une recherche à visée étiologique : le « Manual of operation » rédigé en 1949 constitue le plan le plus proche de ce qu'on appelle aujourd'hui « enquête prospective de cohorte » (Manual of operation, NHI, 1949, Framingham Archives).

Pour le projet d'enquête d'incidence, Meadors pouvait s'inspirer d'études menées par des épidémiologistes de l'*US PHS* quelques années auparavant et qui avaient commencé d'articuler enquête prospective d'incidence et analyse étiologique [38]. Dans ces études, l'unité géographique de la ville avait été retenue. Une enquête prospective d'incidence avait notamment été dirigée par Edgar Sydenstricker dans la ville de Hagerstown (Maryland) entre 1921 et 1924 [39] puis, entre 1938 et 1943 ; sa population fut d'ailleurs considérée comme un « laboratoire de recherche communautaire » [40]. Il nous semble que le choix d'inclure les femmes dans l'enquête de Framingham s'explique aussi principalement par ses débuts comme enquête d'incidence. Comment justifier toutefois qu'elles aient été retenues quand l'étude

devint plus clairement une recherche à visée étiologique ? Dans son livre relatant l'histoire de l'enquête, Thomas Dawber, qui, succédant à Gilcin Meadors, fut le directeur de l'enquête de 1950 à 1965, justifie *a posteriori* ce choix : puisque les femmes étaient justement moins susceptibles de développer ces maladies que les hommes, les inclure permettait de chercher quel est l'éventuel facteur protecteur [18]. Cependant, même dans le manuel de 1949, on ne trouve pas ce motif ainsi explicité ; la question de la différence sexuelle est simplement nommée comme l'une des 28 hypothèses à tester. Puisque le projet était d'obtenir des données rigoureuses sur le taux d'incidence, l'usage de ces données pour tester quelques hypothèses étiologiques par la comparaison des taux d'incidence dans divers sous-groupes et notamment, celle d'une différence sexuelle, semble s'inscrire dans le prolongement de l'enquête d'incidence.

Par ailleurs, quand elle débute en 1947, l'enquête d'incidence n'était qu'un volet d'un programme plus vaste de santé publique du type de « l'enquête communautaire [*community study*] » [41, 42] dont l'objectif est aussi la prévention (éducation à la santé) et le dépistage des maladies d'une communauté particulière. L'unité d'une ville de taille relativement moyenne s'était avérée pertinente et ce, notamment à l'issue d'un programme qui avait été mené sur la tuberculose, précisément à Framingham entre 1917 et 1923 [43-45]. La notion de 'communauté' n'est donc pas à comprendre dans un sens 'ethnique' mais plutôt comme référant à l'unité géographique, sociale et même écologique que constituent une ville et ses habitants. Dès lors, les premiers objectifs combinant une enquête d'incidence et un programme de santé communautaire contribuent à expliquer le choix d'une 'population générale' comme population source. Le premier plan de l'étude, bien que modifié de manière importante en 1949 dans le « Manual of operation », marqua durablement l'enquête.

Le choix de Framingham et la question de la représentativité de sa population

L'État du Massachussetts fut rapidement retenu par Joseph Moutin, directeur du *Bureau of States Services* de l'*US PHS,* pour la réalisation du projet qui serait dès lors mené en collaboration avec le département de santé publique de cet État et avec l'Université de Harvard. Ce choix peut s'expliquer par le fait que des enquêtes de morbidité sur les maladies chroniques avaient déjà été réalisées durant l'entre-deux-guerres dans ce département de santé publique [46] mais aussi à cause de la grande renommée des cardiologues de Boston, tels Samuel Levine et Paul Dudley White, eux-mêmes déjà engagés dans des études cliniques sur quelques centaines de patients et convaincus que les clefs de la prévention viendraient d'études de comparaison de groupes d'individus [47, 48]. Notons d'ailleurs que Paul Dudley

White qui participa au comité de conseil de l'enquête et fut longtemps le conseiller privilégié du *NHI* eut une influence déterminante pour l'investissement fédéral dans cette enquête [18, 49].

Un des critères pour choisir la ville de cet État qui serait retenue fut sa taille. Si Meadors souhaitait pouvoir disposer d'un échantillon d'environ 8000 individus, il fallait une population source d'environ 25 000 personnes. Un cardiologue de l'Université de Harvard, David Rutstein, joua un rôle décisif pour le choix de Framingham parmi trois villes proposées (Beverly, Peabody). Les motifs présentés sont essentiellement pragmatiques. Il souligne le double avantage de cette ville : d'une part, sa proximité avec Boston et donc avec ses cardiologues et ses hôpitaux et, d'autre part, le précédent d'une 'enquête communautaire' sur la tuberculose et, par suite, l'avantage non négligeable de la familiarité de la population avec ce genre d'études (Lettre de Robbins à Meadors, 5 septembre 1947, Framingham Archives).

Mais qu'en fut-il de la question de la représentativité de cette population pour l'enquête d'incidence ? Le choix de l'ensemble de la population d'une ville comme population source évite certains biais mais ne suffit toutefois pas à garantir la représentativité des habitants de cette ville relativement à la population générale américaine. Dans la lettre évoquée précédemment, Robbins rapporte une remarque de Rutstein abordant indirectement cette question : l'intérêt de Framingham est aussi que, malgré sa proximité avec Boston, la ville n'est pas une banlieue-dortoir. On retrouve ces trois arguments dans le premier article publié en 1951 qui présente le plan d'étude de l'enquête. Les auteurs ajoutent le constat que cette ville recèle un centre d'affaires, un centre résidentiel ainsi qu'une aire rurale [37] ; c'est la diversité professionnelle des habitants qui est mise en avant. La représentativité est ici pensée dans les termes de l'échantillonnage raisonné et non dans un sens probabiliste. La partie doit ressembler au tout, dans sa diversité et dans l'ensemble de ses caractéristiques [50]. Toutefois, en 1947, la question de la diversité ethnique ne semble pas avoir été présente au moment du choix de Framingham. Ses habitants sont pourtant essentiellement blancs : il n'y a pratiquement « aucun noir, aucun oriental et la composition de la population blanche n'est pas nécessairement celle des populations blanches d'ailleurs » [18]. Cette question ne fait son apparition que dans l'article de 1951 et probablement sous l'influence d'un statisticien du *NHI*, Felix Moore, qui venait d'être associé à l'étude en 1949 du fait du transfert de cette dernière au *NHI*. Dans cet article, il est en effet précisé que l'idéal serait de réaliser l'étude simultanément dans plusieurs communautés géographiques des États-Unis « de sorte que divers groupes ethniques et raciaux soient représentés » [37].

Toutefois, quand l'objectif de recherche étiologique devient prioritaire après 1949, il s'agit alors plutôt d'analyser les corrélations de variables individuelles à l'intérieur de la population d'étude que d'obtenir une mesure de référence pour le taux d'incidence de ces maladies dans la population américaine. Dès lors, la contrainte de représentativité de la population d'étude est moins importante ; l'essentiel est qu'elle ait un nombre suffisant de sujets dotés des caractéristiques étudiées pour la comparaison. En outre, la variance d'une caractéristique donnée pouvait s'avérer plus grande et plus intéressante au sein même de la population de Framingham qu'entre différentes villes. Voici ce que Dawber, Meadors et Moore écrivent en 1951 : « Cette limitation dans la couverture géographique restreint clairement la généralité des conclusions qui peuvent être atteintes. Il y a, cependant, un fondement raisonnable à la conviction selon laquelle la distribution de l'artériosclérose et de l'hypertension dans la race blanche aux États-Unis est telle que la variance dans la communauté est bien plus importante que la variance entre les communautés (…) » [37]. Reste cependant que l'insuffisante diversité ethnique demeure une limite pour l'étude étiologique elle-même et a soulevé des critiques par la suite, les différences ethniques pouvant jouer un rôle important dans le risque cardiovasculaire [51].

LA CONSTITUTION DE LA COHORTE

Intéressons-nous désormais à la manière dont fut constituée la cohorte qui, initialement prévue pour être de 8000 individus, comporta finalement 5209 individus. A cause du suivi dont elle devait être l'objet, les investigateurs durent concilier diverses contraintes parfois contradictoires.

Tranches d'âge et nombre d'individus

Gilcin Meadors retint la tranche d'âge entre 30 et 60 ans. Puisqu'il s'agissait d'obtenir une mesure du taux d'incidence des maladies coronariennes, la cohorte devait permettre d'obtenir rapidement un grand nombre de nouveaux cas. Toutefois, pour l'étude étiologique de ces maladies au développement lent et progressif et dont les origines peuvent remonter à l'enfance, on pouvait imaginer qu'une tranche d'âge qui s'élargisse aux individus de moins de 30 ans soit utile. S'il fut reconnu, dans l'article de 1951 précédemment évoqué, que l'idéal aurait été de « suivre une cohorte d'individus de la naissance à la mort » [37], cela n'était cependant pas apparu comme un objectif réalisable. La plus grande mobilité des moins de 30 ans fut notamment évoquée comme argument pour ne pas élargir la cohorte à ces individus.

Pour ce qui concerne le nombre de sujets nécessaires dans cette tranche d'âge, les 8000 individus évoqués par Meadors dans la première mouture du projet en juillet 1947 représentent un tiers environ des habitants. On trouve peu d'explication pour ce nombre qui dans le manuel de 1949 est réduit de quelques milliers (environ 6000). Des considérations sur le nombre de cas nécessaires pour qu'une évaluation du taux d'incidence et une comparaison des taux dans divers groupes soient possibles avaient toutefois été faites (Meadors, 4 Juin 1948, Statistical Planning for Heart Disease Epidemiology Study in Framingham, Framingham Archives). L'expérience précédente de l'enquête sur la tuberculose a pu être aussi influente : un tiers des résidents avait aussi été retenu [52]. Des critères plus pragmatiques ont aussi probablement compté. Il fallait en effet que l'ensemble des individus de la cohorte puisse être examiné dans un intervalle de temps suffisamment court pour que l'évolution des facteurs ou l'apparition de nouveaux symptômes soient dépistés. Le rythme bi-annuel fut retenu. On envisagea un rythme d'examen de 300 individus par mois environ, à raison de 30 minutes par personne.

Principe de recrutement

Pour la sélection des individus, un mode aléatoire d'échantillonnage aurait pu être envisageable. Or la cohorte fut d'abord constituée à partir de volontaires, avec la famille comme unité statistique. Pour Meadors, quand les examens cliniques débutèrent en 1948, il semblait suffire que la population source fût « une population 'normale' et non sélectionnée ». Il ne s'inquiéta pas des biais que le principe du volontariat pouvait introduire. On peut pourtant aisément imaginer que les volontaires fussent uniquement les individus les plus inquiets et déjà les plus susceptibles d'être atteints de ces maladies. Plutôt que guidé par des principes statistiques, Meadors sembla surtout répondre à des contraintes pratiques et s'inspirer là encore de l'enquête de Framingham sur la tuberculose aussi menée à partir d'individus volontaires.

En réalité, la randomisation comme condition de validité de l'inférence statistique ne faisait alors que commencer à s'imposer dans les enquêtes statistiques. La méthode avait été introduite dans le cadre des travaux de Ronald Fisher puis de Egon Pearson et Jerzy Neyman dans les années 1930 [53]. Du côté de la statistique médicale, la randomisation fut utilisée dans les années 1940 à l'occasion des premiers essais cliniques [20]. Elle s'introduisit aussi progressivement dans les sciences sociales et dans la pratique des statisticiens du gouvernement fédéral américain [50]. Felix Moore s'était très certainement familiarisé avec cette méthode probabiliste avant d'arriver au *NHI*. Il proposa de modifier la cohorte et de

s'appuyer sur la liste des résidents de la ville : un individu sur trois serait retenu à partir de la liste alphabétique. Dès lors, l'individu devenait l'unité statistique sans qu'il ne fût tenu compte des familles. Précisons qu'un échantillonnage à partir des registres de recensement était facilité par le fait que la ville de Framingham et l'État du Massachusetts disposaient de registres exhaustifs et parmi les plus rigoureusement tenus.

Mais cette contrainte probabiliste de sélection aléatoire eut toutefois à composer avec une autre contrainte forte pour la faisabilité de l'enquête de cohorte dont nous reparlerons : la participation de la population. C'est probablement à cause d'une mauvaise évaluation des pertes dans l'échantillon sollicité (68,8% seulement de l'échantillon appelé accepta de participer) que Moore dut finalement rappeler l'ensemble des volontaires éligibles ; notons d'ailleurs qu'avec un tel taux de participation, les taux de prévalence obtenus ne pouvaient être qu'approximatifs. Mais il semble surtout qu'il apparut vite risqué, du point de vue de l'adhésion de la population au projet, de ne pas inclure aussi les volontaires qui avaient commencé de participer à l'étude depuis 1947. C'est sous la pression du Comité exécutif de l'enquête, comité constitué de représentants de la ville de Framingham, que Moore serait finalement revenu à l'unité d'étude de la famille plutôt qu'à celle de l'individu [37]. Aussi la cohorte finale fut-elle constituée du mélange de ce tiers randomisé à partir de la liste des résidents avec une partie de la précédente cohorte de volontaires (740). Les volontaires constituèrent 14% de la cohorte finale [12]. Moore eut le souci de s'assurer que l'inclusion de la population de volontaires n'introduisit pas trop de biais (Moore à Van Slyke, 26 août 1949, *Memorandum*, Framingham Archives).

Sains et malades

Il y a deux temps pour la constitution d'une cohorte en vue d'une enquête d'incidence. Puisqu'on ne suit en principe que les individus qui n'ont pas encore la maladie étudiée, une étude sur un premier échantillon qui consiste en une 'recherche de cas [*case-finding*]' est donc nécessaire pour exclure ces derniers de la future cohorte. L'examen clinique initial qui dura de fin 1947 à 1952 fut le plus complet. Mais contrairement à ce qui était prévu, l'ensemble des individus furent finalement suivis, les sains comme les malades. Comment expliquer cette décision qui alourdissait le coût de l'étude ?

Deux difficultés majeures s'étaient posées dans l'exclusion des individus déjà malades, l'une d'ordre diagnostique, l'autre d'ordre pratique. La première est liée aux nombreux 'cas frontières'. Les deux catégories nosologiques retenues pour l'étude étaient alors les « maladies athérosclérotiques et hypertensives ». À l'issue des premiers examens,

malgré toutes les précautions prises avec, d'une part, la définition préalable de critères diagnostiques précis et de leur interprétation (*Manual of operation*) et, d'autre part, un examen clinique détaillé réalisé par deux médecins, le diagnostic de nombreux sujets demeuraient difficile à établir. Pour ce qui concerne la « maladie athérosclérotique », des désaccords concernaient notamment les diagnostics d'angine de poitrine, l'un des trois grands syndromes retenus pour la définir avec l'infarctus du myocarde et la mort subite. Il fut décidé de ne pas exclure les cas douteux de la cohorte. Une telle décision a une influence considérable sur la valeur des taux de prévalence et d'incidence : si cela conduit à un moindre taux de prévalence, les taux d'incidence se trouvent quant à eux augmentés. C'est toutefois le suivi sur le long terme qui, précisément, devait aider à la catégorisation de ces cas problématiques [18, 54]. Par ailleurs, le problème de la « maladie hypertensive » est qu'elle a un double statut : elle est à la fois *end point* (ce qui requiert que les hypertendus soient exclus) et hypothèse étiologique (ce qui requiert qu'ils soient inclus). De plus, la définition de la pression artérielle normale faisait débat. Des études statistiques sur sa distribution dans une population générale, qui succédaient à celles des assureurs [55, 56], montraient qu'elle est unimodale et qu'il n'existe peut-être aucune division naturelle entre normotension et hypertension [57]. Pour toutes ces raisons, les investigateurs décidèrent finalement de ne pas exclure de la cohorte ceux qui avaient été initialement classés comme « hypertendus » [18].

La deuxième difficulté concernait l'exclusion des individus alors identifiés comme malades : une telle exclusion risquait là aussi de mettre en péril la participation de la population à l'étude. Voici ce que les investigateurs écrivirent à ce sujet en 1957 : « Le plan initial de l'enquête de Framingham prévoyait de ne suivre que les personnes de l'échantillon qui ne seraient pas atteintes de la maladie athérosclérotique et de la maladie hypertensive. Il fut vite évident que restreindre l'attention sur ce groupe, identifié au mieux avec difficulté, conduirait à la perte d'informations très utiles pour l'histoire de la maladie coronarienne. De plus, il apparut que l'exclusion des personnes malades après le premier examen aurait affaibli la relation de la médecine clinique avec la communauté » [54].

Ainsi, d'une part, même les cas les plus aigus, bien qu'exclus de la cohorte, firent l'objet d'un suivi clinique et d'autre part, les cas pour lesquels le diagnostic était problématique furent inclus dans la cohorte. Cela introduisait une grande souplesse dans la classification, laissant la possibilité que le diagnostic fût infirmé ou confirmé sur le long terme. Ce suivi de l'ensemble du premier échantillon contribua à faire de la population d'étude de Framingham un « laboratoire statistique » particulièrement précieux. Pour Thomas Dawber, ajouté à la grande précision et l'uniformité de l'examen clinique des sujets, un tel

suivi constitue le second grand avantage des données issues de la cohorte de Framingham relativement aux autres études [18].

Suivi et participation de la population

La longueur du suivi qui, initialement prévue pour 10 ans, passa ensuite à 20 ans et bien au-delà puisque l'étude se poursuit encore aujourd'hui, soulevait des difficultés nouvelles aux statisticiens. En effet, le risque de perdus de vue est l'une des difficultés majeures de l'enquête prospective de cohorte : il augmente proportionnellement à la durée de l'étude. La pertinence et la validité des comparaisons repose sur la stabilité de la cohorte. Cette longueur du suivi peut aussi mettre en péril la participation de la population dont l'adhésion au projet est déterminante. Tous les deux ans, les individus de la cohorte doivent se présenter à la clinique pour un examen détaillé.

L'enquête de Framingham fait aujourd'hui figure d'exception par la qualité de la participation de sa population d'étude. Il y eut tout un travail de la part de représentants de la ville et d'associations locales pour sensibiliser l'ensemble des habitants de Framingham à l'intérêt de l'étude et faciliter la participation [12], selon des modalités très comparables à ce qui s'était passé pour le programme sur la tuberculose. C'est aussi très certainement parce que la population a d'abord été le *sujet* d'un projet communautaire de santé publique avant de devenir davantage l'*objet* d'une recherche étiologique qu'elle adhéra à ce projet. Ainsi, paradoxalement, si les adaptations successives dans la constitution de la population font que l'étude est loin de correspondre aux standards actuels du protocole de l'enquête de cohorte, tous les éléments hérités de ses débuts tâtonnants jouèrent très certainement en faveur de la participation de la population.

Ainsi, le choix de la population générale d'une ville, le suivi des sains et des malades et la bonne participation de la population expliquent la richesse et la qualité des données recueillies à partir de la cohorte de Framingham depuis 1947. En 1971, une deuxième cohorte (Offspring Study) fut constituée d'un échantillon de 5135 individus à partir de la progéniture de la cohorte originelle. La cohorte initiale se réduisait et vieillissait de manière trop importante. Sur cette deuxième cohorte, il n'y avait, en 1998, que 20 individus perdus de vue. En 1998 toujours, sur les 5209 individus de la cohorte originelle, il restait environ 1095 individus vivants. Une troisième (Generation III cohort) réalisée à partir des enfants de la seconde cohorte, en cours de constitution a pour finalité l'étude des facteurs génétiques. L'enquête de Framingham constitue l'une des sources de données les plus abondantes et les

plus rigoureuses sur trois générations successives, une mine pour l'épidémiologie génétique des maladies coronariennes, mais aussi pour de nombreuses autres maladies. Dès les années 1960, les données de la cohorte de Framingham furent sollicitées par d'autres épidémiologistes dans le but de mener des comparaisons et des analyses ou de tester des hypothèses. Aujourd'hui, les investigateurs de Framingham collaborent avec d'autres chercheurs pour l'analyse de maladies comme l'ostéoporose, la démence, le cancer du poumon, l'arthrite, le diabète, les maladies oculaires, ainsi que le profil génétique de maladies communes. On comprend mieux dès lors que Thomas Dawber ait parlé de la cohorte de Framingham comme d'un « laboratoire populationnel [*population laboratory*] » [18]. Toutefois, les nouvelles cohortes, à la différence de la première, sont désormais constituées d'individus qui prennent des médicaments contre leurs facteurs de risque. Il est alors bien plus difficile d'obtenir des données non biaisées pour l'analyse. L'étude épidémiologique d'observation se trouve ici confrontée à des difficultés nouvelles.

L'ANALYSE DES DONNÉES : L'ÉMERGENCE DE L'APPROCHE « FACTEURS DE RISQUE »

Un deuxième point fort de l'enquête réside dans l'adaptation de modèles mathématiques multivariés pour l'analyse des données épidémiologiques. La corrélation multiple était déjà utilisée en anthropologie mais c'est à l'occasion de l'analyse des données de Framingham que la modélisation multivariée commença de devenir un outil fondamental de l'épidémiologiste pour appréhender la multifactorialité des maladies. Les cliniciens de l'enquête qui dirigeaient l'étude et assuraient les examens dans une clinique à Framingham n'étaient pas formés en statistique ni même en épidémiologie. Les données recueillies et mises en fiche étaient envoyées au bureau de biométrie du *NHI* à Bethesda (Maryland). Aux *National Institutes of Health* (*NIH*), depuis 1948, le statisticien-sociologue Harold Dorn, qui avait suivi les enseignements à Londres d'Egon Pearson et de Ronald Fischer [58], avait réuni autour de lui des statisticiens particulièrement compétents qui contribuèrent grandement à l'adaptation des techniques de la nouvelle statistique inférentielle en médecine et en épidémiologie (Jerome Cornfield, Nathan Mantel, William Haenszel, etc.) [59]. Harold Dorn joua d'ailleurs un rôle majeur dans la théorisation de l'étude prospective [60, 61]. Sans la compétence de ce groupe de statisticiens, les données de l'enquête n'auraient probablement pas été analysées avec une telle rigueur. Toutefois, cette répartition des tâches entre cliniciens et statisticiens souleva des difficultés, ne serait-ce qu'à cause de la distance géographique

entre Framingham et Bethesda et il fut nécessaire d'articuler langages, conceptions et pratiques. L'« approche facteurs de risque de la maladie », expression que nous empruntons à Robert Aronowitz [62], est le fruit de ce travail d'articulation entre clinique et statistique [13].

DE LA METHODE CLASSIFICATOIRE À L'ANALYSE MULTIVARIEE

Les premières analyses : les tableaux de contingence (1957-1961)

Pour l'analyse de l'association entre une variable d'exposition et une maladie, la méthode la plus simple alors utilisée en épidémiologie repose sur la classification des personnes en autant de sous-groupes que de variables d'exposition étudiées et un calcul du taux d'incidence pour chacun de ces sous-groupes. Le tableau de contingence permet de les comparer : en 1957, pour la pression artérielle par exemple, apparut très nettement la grande différence entre le taux d'incidence du sous-groupe normotendu (26 pour 1000) et celle du sous-groupe d'individus ayant une hypertension certaine (98 pour 1000). Précisons que ces premières analyses de corrélation ne furent pas établies à partir de la cohorte entière mais seulement d'un sous-groupe de la cohorte : des hommes ayant entre 45 et 62 ans (898 individus) [54]. Il fallut attendre 1961 pour qu'il y eût suffisamment de nouveaux cas aussi chez les femmes et qu'il soit possible de les inclure et d'élargir la tranche d'âge à 40-59 ans [63]. Parmi les 28 variables d'exposition ou hypothèses qui avaient été retenues et décrites en détail dans le manuel de 1949, une poignée d'entre elles émergea rapidement comme fortement associée à un surcroît de risque de morbi-mortalité coronarienne : le sexe masculin, l'âge, l'hypertension, l'hypercholestérolémie, le surpoids, l'hypertrophie ventriculaire gauche. Puis le tabagisme fut rapidement ajouté [64, 65]. Il apparut aussi qu'aucune hypothèse ne l'emportait plus nettement sur les autres. Dans l'article qui livre les premiers résultats de l'enquête en 1957, il n'est pas donné de mesure quantitative de l'association mais les investigateurs se sont cependant assurés de la signification statistique des différences observées entre les taux d'incidence [54]. En 1961, disposant de davantage de données, l'association est quantifiée par le calcul du risque relatif [63]. Ces risques relatifs furent standardisés par leur comparaison avec ceux calculés à partir de taux d'incidence attendus. La première apparition dans le domaine médical de l'expression « risk factor » déjà utilisée dans les milieux des assurances et désignant les attributs associés à une augmentation du taux d'incidence daterait de l'article de 1961 [66, 67]. En réalité, il nous semble surtout que les cliniciens de l'enquête jouèrent un rôle particulièrement important dans la diffusion de la notion et de l'approche de la maladie qu'elle induit auprès de leurs confrères [68, 69].

En 1957, par le moyen d'un tableau un peu plus complexe réalisant plusieurs combinaisons, les investigateurs vérifièrent que pour chaque facteur, le taux d'incidence augmentait quel que soit le niveau des autres facteurs (*tableau I*). L'hypertension comme l'hypercholestérolémie s'avéraient contribuer chacun de manière indépendante au risque. En revanche, le facteur surpoids perdait de son importance. Un troisième type de tableau fut réalisé pour évaluer l'effet conjoint des variables de risque. Diverses catégories furent établies selon que les individus avaient une variable élevée, deux et enfin trois (*tableau II*). Il apparut alors clairement que le taux d'incidence de la maladie coronarienne augmente avec le nombre et l'élévation du niveau des facteurs présents et que l'augmentation est progressive [54]. Dans la publication de 1961, à ce type de tableaux, les investigateurs ajoutèrent une représentation graphique de l'effet conjoint du cholestérol et de la pression artérielle qui mettait encore mieux en évidence l'apparence progressive de cette augmentation [63].

Les limites de la méthode classificatoire et les premières équations (1962-1967)

L'identification de facteurs de risque et l'analyse de corrélation dans le cadre d'une procédure classificatoire se trouvaient confrontées à trois principales limites que les statisticiens qui succédèrent à Felix Moore dans la responsabilité de l'analyse de l'enquête mirent en évidence. Ce sont ces limites qui conduisirent Jerome Cornfield à justifier le recours à un modèle mathématique dans un article publié en 1962 [70]. Premièrement, une procédure classificatoire requiert que les variables quantitatives soient transformées en variables qualitatives. Les investigateurs, conscients du caractère relativement arbitraire de la frontière entre le normal et le pathologique pour la pression artérielle ou le taux de cholestérol par exemple, avaient opté pour une première solution consistant à multiplier les groupes et à établir des catégories frontières (voir *tableau I*). Mais une telle solution induit une perte importante d'information qui peut nuire au repérage de seuils éventuels ou « valeurs critiques » dans la relation de risque. Elle s'avère d'autant plus problématique pour l'épidémiologie des maladies coronariennes que les principaux facteurs de risque, le taux de cholestérol, la pression artérielle et le poids, sont des variables quantitatives continues. Deuxièmement, avec un tableau, on n'a pas d'information quantitative sur la manière dont les effets des variables se combinent dans le risque de la maladie. L'élimination d'éventuels effets confondants d'autres variables n'est pas assurée. Cornfield souligne qu'il importe aussi de se demander s'il y a « potentialisation » des effets des variables dans le risque et de pouvoir la quantifier. C'est une manière de désigner ce qu'on appelle aujourd'hui « interaction ». Troisièmement, ce type d'analyse est possible avec deux ou trois variables

mais au-delà, les tableaux deviennent extrêmement complexes. Dans un article publié en 1967, ce dernier problème est bien exprimé : « La méthode analytique traditionnelle de l'épidémiologiste, la classification croisée multiple [*multiple cross-classification*], devient rapidement impraticable à mesure que le nombre de variables à étudier augmente. Donc, si 10 variables sont observées et que chaque variable doit être étudiée à seulement trois niveaux, par exemple les niveaux de cholestérol à moins de 225 mg/100 ml, 225-274, et 275 et plus, il y aurait 59 049 cellules dans la classification croisée multiple » [71].

Ainsi, aux yeux de Cornfield, dans le cadre d'une étude où l'accumulation des données est lente, le modèle mathématique qui résume les observations dans un petit nombre de paramètres disponibles et offre les moyens de répondre quantitativement aux questions posées est particulièrement avantageux. Les valeurs des variables quantitatives peuvent être introduites dans l'équation sans qu'il soit nécessaire de les catégoriser préalablement. Le premier modèle utilisé par Cornfield dans l'article de 1962 est une adaptation de la fonction discriminante mise au point par Ronald Fisher, une fonction discriminante linéaire, qu'il applique à l'analyse du rôle conjoint de la pression artérielle et du niveau de cholestérol. Il avait déjà utilisé ce modèle dans le cadre d'études expérimentales en laboratoire pour évaluer les effets conjoints et indépendants des granulocytes et des lymphocytes sur la survie de souris irradiées en 1961 [72]. Son innovation dans l'article de 1962 est de justifier et défendre l'intérêt de son utilisation pour des données épidémiologiques d'observation. L'équation permit de confirmer la continuité de la relation de risque pour la pression artérielle et pour le cholestérol ainsi que leur contribution indépendante au risque en en donnant une valeur quantitative. Il apparaissait aussi que l'action conjointe de ces deux facteurs est de nature « multiplicative ».

Un second modèle fut utilisé en 1967 suite à un travail en collaboration avec une autre statisticienne du *NHI*, Jeanne Truett. Ils utilisèrent une équation logistique multiple qui permit d'appliquer la fonction mathématique à un plus grand nombre de variables (sept : âge, cholestérol, pression artérielle, poids, taux d'hémoglobine, consommation de tabac, présence d'une anormalité à l'électrocardiogramme) [71]. L'analyse confirma les précédents résultats de Cornfield. Par la suite, une version modifiée du modèle de Walker et Duncan fut utilisée [73]. Les coefficients de leur équation sont plus facilement calculables et évitent des hypothèses coûteuses concernant l'allure normale de la distribution des variables, requises par le modèle de Truett et Cornfield. Ces analyses quantitatives de la corrélation multiple légitimèrent d'un point de vue statistique, l'attribution du statut de 'facteur de risque' à ces diverses variables.

DE L'USAGE ANALYTIQUE A L'USAGE PREDICTIF DES EQUATIONS DE RISQUE

Aussi les premiers usages des équations multivariées eurent-ils une finalité essentiellement *analytique* : il s'agissait de simplifier et faciliter l'analyse de corrélation. L'usage dit *synthétique,* qui permet de tester la validité de l'équation et de ses coefficients sur d'autres données d'observation que celles à partir desquelles l'équation a été constituée, a aussi une visée d'analyse. Mais rapidement, les cliniciens et surtout, l'*American Health Association,* sollicitèrent un usage *synthétique* de ces équations dans un but pratique de prédiction du risque d'un individu particulier. Dans ce cas, l'équation, en combinant l'information à partir d'un ensemble donné de facteurs de risque, est utilisée pour estimer la probabilité que des individus qui ont tel ou tel facteur développent une maladie coronarienne dans une période fixée de temps. Avant même que de telles équations soient constituées, des tables de risque à destination des médecins avaient été établies dès 1971 pour aider à la prédiction du risque individuel [74, 75].

Mais en dépit de la continuité apparente des usages *analytique* et *synthétique* de ces équations, il y a un véritable saut de l'un à l'autre. Un autre statisticien du *NHI,* Tavia Gordon, insista tout particulièrement sur la spécificité des problèmes qui se posent pour un usage synthétique et prédictif des équations de Framingham dans un but préventif [74]. L'usage synthétique implique une forme de généralisation à partir de l'expérience de Framingham. Il était tout d'abord nécessaire de s'assurer de « l'aptitude prédictive » de l'équation. Elle fut testée sur un ensemble compilé des données des autres études prospectives sur les maladies coronariennes initiées dans d'autres États américains : le « pooling project » mené sous la direction de Felix Moore. Les résultats en 1976, après un travail important de standardisation des données, furent positifs et encouragèrent la généralisation de l'équation [76]. Il convient aussi de souligner qu'avec une finalité de prédiction plutôt que d'analyse, les facteurs à retenir pour l'équation ne sont pas nécessairement les mêmes. La valeur prédictive d'un facteur n'est pas toujours identique à son importance étiologique. Par exemple, bien que le poids soit un facteur étiologique important, il se révèle peu utile pour l'équation de prédiction. Par ailleurs, ces facteurs doivent être aisément et rapidement mesurables par le médecin. C'est en 1976 qu'est publiée dans l'*American Journal of Cardiology* la première fonction prédictive de risque visant à faciliter l'identification des individus à haut risque de différentes maladies coronariennes [77].

Ainsi, dans le cadre de l'enquête de Framingham, l'« approche facteurs de risque » des maladies se constitue comme une sorte de mathématisation de la multifactorialité des

maladies selon deux orientations, explicative et prédictive, qui, quoique souvent confondues, sont bien différentes.

CONCLUSION

S'il convient clairement de relativiser le statut exemplaire de son plan d'étude, l'enquête de Framingham a bien joué un rôle fondateur à cause de sa contribution à l'émergence du rôle nouveau de l'enquête de cohorte dans la recherche étiologique. Paradoxalement, le tâtonnement des débuts a d'une certaine façon contribué à faire de sa cohorte une « population laboratoire ». Par ailleurs, c'est à l'occasion des premières analyses de ses données que furent introduits et adaptés les modèles multivariés aujourd'hui si centraux en épidémiologie analytique. Ces modèles rendirent possible une modélisation performante de la multifactorialité, étape essentielle dans la constitution du concept épidémiologique de 'facteur de risque'.

L'histoire de l'enquête de Framingham révèle des continuités avec l'épidémiologie qui précède la Seconde Guerre mondiale : le plan d'étude de l'enquête prospective à visée étiologique y apparaît comme un prolongement des enquêtes d'incidence développées dès le premier tiers du 20^{e} siècle. Elle montre aussi qu'au lendemain de la Seconde Guerre mondiale, les standards et les règles de l'enquête prospective de cohorte se mettent peu à peu en place dans une négociation permanente entre cliniciens et statisticiens. La statistique inférentielle est progressivement intégrée à la méthodologie de l'enquête et aux outils d'analyse des données, dans une recherche constante de compromis entre exigence d'objectivité et faisabilité. Dès lors, les études étiologiques d'observation en épidémiologie sont une sorte de compromis entre scientificité et pragmatisme. Ce compromis est certainement l'enjeu central des études épidémiologiques qui ont l'avantage de permettre le recueil d'informations précieuses sur des êtres humains dans leur milieu de vie.

REMERCIEMENTS

Pour la réalisation de cet article, nous tenons à remercier Joël Coste et Alfred Spira. Le travail de recherche sur l'histoire de l'enquête de Framingham est une partie d'une réflexion menée au sein d'une thèse en philosophie et histoire des sciences intitulée « Epidémiologie des facteurs de risque : genèse d'une nouvelle approche de la maladie », dirigée par le Professeur Jean Gayon et soutenue en décembre 2006. Le travail sur les archives de l'enquête de

Framingham a été effectué grâce au soutien, d'une part, de la Pisano Grant, de l'Office of NIH History où nous remercions plus particulièrement Victoria Harden et à l'accueil du National Heart, Lung and Blood Institute où nous remercions Paul Sorlie et, d'autre part, grâce au soutien de J.P. Gaudillière, I. Lowy et L. Berlivet, dans le cadre d'un programme de recherche « risque et santé » CNRS/CERMES. Nous remercions aussi tout particulièrement le Professeur Pierre Corvol pour ses précieux encouragements.

REMARQUE

Les archives non publiées auxquelles nous faisons référence dans le texte sont issues de la source suivante : Archives du National Heart, Lung and Blood Institute, Framingham Heart Study Papers, Bethesda, Maryland : Correspondance, Mémorandums, Rapports d'activité, Monographies.

REPRODUCTION

Les deux tableaux sont reproduits avec la permission de l'American Public Health Association

RÉFÉRENCES

1. Paneth N, Susser E, Susser M. Origins and early development of the case-control study: Part 1, Early evolution. Soz Praventivmed 2002;47:282-8.
2. Paneth N, Susser E, Susser M. Origins and early development of the case-control study: Part 2, The case-control study from Lane-Claypon to 1950. Soz Praventivmed 2002;47:359-65.
3. Gordon J. The twentieth century – yesterday, today, and tomorrow In: Top F, ed. The History of American Epidemiology. Saint Louis: Mosby, 1952:114-67.
4. MacMahon B. Epidemiologic methods. 1st ed. Boston,: Little Brown, 1960.
5. Rothman K. Epidemiology: an introduction. Oxford: Oxford University Press, 2002.
6. Susser M. Epidemiology in the United States after World War II: the evolution of technique. Epidemiol Rev 1985;7:147-77.
7. Doll R. Cohort studies: history of the method. I. Prospective cohort studies. Soz Praventivmed 2001;46:75-86.
8. Doll R. Cohort studies: history of the method. II. Retrospective cohort studies. Soz Praventivmed 2001;46:152-60.
9. Liddell FD. The development of cohort studies in epidemiology: a review. J Clin Epidemiol 1988;41:1217-37.
10. Morabia A, Guthold R. Wilhelm Weinberg's 1913 Large Retrospective Cohort Study: a rediscovery. Am J Epidemiol 2007;165:727-33.
11. Comstock GW. Cohort analysis: W.H. Frost's contributions to the epidemiology of tuberculosis and chronic disease. Soz Praventivmed 2001;46:7-12.
12. Oppenheimer GM. Becoming the Framingham Study 1947-1950. Am J Public Health 2005;95:602-10.

13. Giroux E. Epidémiologie des facteurs de risque: genèse d'une nouvelle approche de la maladie, Thèse de doctorat en philosophie de la médecine. Paris: Université de Paris 1 Panthéon Sorbonne, 2006.
14. Oppenheimer GM. Profiling risk: the emergence of coronary heart disease epidemiology in the United States (1947-70). Int J Epidemiol 2006;35:720-30.
15. Aronowitz R. La construction sociale des facteurs de risque des maladies coronariennes. In: Aronowitz R, ed. Les maladies ont-elles un sens? Paris: Synthélabo, 1999:223-89.
16. Rothstein WG. Public health and the risk factor : a history of an uneven medical revolution. Rochester, NY: University of Rochester Press, 2003.
17. Levy D, Brink S. A change of heart : how the Framingham heart study helped unravel the mysteries of cardiovascular disease. 1st ed. New York: Knopf, 2005.
18. Dawber TR. The Framingham study : the epidemiology of atherosclerotic disease. Cambridge, Mass.: Harvard University Press, 1980.
19. Fox DM. Health policy and changing epidemiology in the United States: chronic disease in the twentieth century. Trans Stud Coll Physicians Phila 1988;10:11-31.
20. Marks H. La médecine des preuves. Histoire et anthropologie des essais cliniques [1900-1990]. Le Plessis-Robinson: Institut Synthélabo, 1999.
21. Moriyama IM, Woolsey TD. Statistical studies of heart disease. IX. Race and sex differences in the trend of mortality from the major cardiovascular-renal diseases. Public Health Rep 1951;66:355-68.
22. Herrick JB. Clinical features of the coronary arteries. J Am Med Assoc 1912;59:2015-20.
23. Herrick JB, Nuzum FR. Angina pectoris, clinical experience with two hundred cases. J Am Med Assoc 1918;70:67-70.
24. Anitschkov N, Chalatow SS. Über experimentelle Cholesterinsteatose und ihre Bedeutung für die Entstehung einiger pathologischer Prozesse. Zentralblatt für allgemeine Pathologie und pathologische Anatomie 1913;24:1-9.
25. Keys A. Nutrition in relation to the aetiology and courses of diseases. Journal of the American Dietetic Association 1948;24:281-5.
26. Paul JR. Clinical epidemiology. J Clin Invest 1938;17:539-41.
27. Gordon JE. Medical ecology and the public health. Am J Med Sci 1958;235:337-59.
28. Epstein FH. An epidemiological study in a total community: the Tecumseh project. Med Bull (Ann Arbor) 1960;26:307-14.
29. Francis T. Aspects of the Tecumseh study. Public Health Rep 1961;76:963-5.
30. Keys A, Taylor HL, Blackburn H, Brozek J, Anderson JT, Simonson E. Coronary heart disease among Minnesota Business and professional men followed fifteen years. Circulation 1963;28:381-95.
31. Doyle JT, Heslin AS, Hilleboe HE, Formel PF, Korns RF. A prospective study of degenerative cardiovascular disease in Albany: report of three years' experience. I. Ischemic heart disease. Am J Public Health Nations Health 1957;47:25-32.
32. Chapman JM, Goerke LS, Dixon W, Loveland DB, Phillips E. Measuring the risk of coronary heart disease in adult population groups. The clinical status of a population group in Los Angeles under observation for two to three years. Am J Public Health Nations Health 1957;47:33-42.
33. Dyer AR, Stamler J, Berkson DM, Lindberg HA. Relationship of relative weight and body mass index to 14-year mortality in the Chicago Peoples Gas Company study. J Chronic Dis 1975;28:109-23.
34. Paul O, Lepper MH, Phelan WH, Dupertuis GW, Macmillan A, McKean H, *et al.* A Longitudinal Study of Coronary Heart Disease. Circulation 1963;28:20-31.

35. Morris JN, Heady JA, Barley RG. Coronary heart disease in medical practitioners. Br Med J 1952;1:503-20.
36. Morris JN, Raffle PA. Coronary heart disease in transport workers; a progress report. Br J Ind Med 1954;11:260-4.
37. Dawber TR, Meadors GF, Moore FE, Jr. Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health Nations Health 1951;41:279-81.
38. Goldberger J. A study of the relation of diet to pellagra incidence in seven textile-mill communities of South Carolina in 1916. Public Health Rep 1920;35:648-713.
39. Sydenstricker E. A study of illness in a general population group. Hagerstown Morbidity Studies No. I : the methods of study and general results. Public Health Rep 1926;41:2069-88.
40. Turner VB. Hagerstown Health Studies: An Annotated Bibliography. Washington Federal Security Agency, Public Health Service Publication no. 148, 1952.
41. Bulmer M, Bales K, Sklar KK. The Social survey in historical perspective, 1880-1940. Cambridge ; New York: Cambridge University Press, 1991.
42. Converse JM. Survey research in the United States : roots and emergence 1890-1960. Berkeley: University of California Press, 1987.
43. Framingham Community Health and Tuberculosis Demonstration. Framingham Monograph No. 10: Final summary report: 1917-1923 Inclusive, General series IV. Framingham, Mass., 1924.
44. Comstock GW. Commentary: the first Framingham Study--a pioneer in community-based participatory research. Int J Epidemiol 2005;34:1188-90.
45. Susser M, Stein Z. Commentary: Donald Budd Armstrong (1886-1968)--pioneering tuberculosis prevention in general practice. Int J Epidemiol 2005;34:1191-3.
46. Bigelow GH, Lombard HL. Cancer and other chronic diseases in Massachusetts. Boston New York: Houghton Mifflin Co., 1933.
47. Gertler MM, Driskell MM, Bland EF, Garn SM, Lerman J, Levine SA, *et al.* Clinical aspects of coronary heart disease; an analysis of 100 cases in patients 23 to 40 years of age with myocardial infarction. J Am Med Assoc 1951;146:1291-5.
48. Gertler MM, Garn SM, White PD. Young candidates for coronary heart disease. J Am Med Assoc 1951;147:621-5.
49. Paul O. Take heart : the life and prescription for living of Dr. Paul Dudley White. Boston: Distributed by the Harvard University Press for the Francis A. Countway Library of Medicine; 1986.
50. Desrosières A. La politique des grands nombres : histoire de la raison statistique. Paris: Editions La Découverte, 1993.
51. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, *et al.* Multi-Ethnic Study of Atherosclerosis: Objectives and Design. Am J Epidemiol 2002;156:871-81.
52. Armstrong DB. The Framingham Health and Tuberculosis Demonstration. Am J Public Health 1917;7:318-22.
53. Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L. The empire of chance : how probability changed science and everyday life. Reprinted ed. Cambridge: Cambridge University Press, 1991.
54. Dawber TR, Moore FE, Mann GV. Coronary heart disease in the Framingham study. Am J Public Health Nations Health 1957;47:4-24.
55. Fisher JW. The diagnostic value of the sphygmomanometer in examinations for life insurance. JAMA 1914;63:1752-4.
56. Actuarial Society of America. Blood pressure study, 1939. New York: Association of Life Insurance Medical Directors of America, 1940.

57. Pickering GW, Roberts JA, Sowry GS. The aetiology of essential hypertension. 3. The effect of correcting for arm circumference on the growth rate of arterial pressure with age. Clin Sci (Lond) 1954;13:267-71.
58. Greenhouse S. Dorn, Harold Fred. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics. New York: J. Wiley, 1998:955-9.
59. Greenhouse SW. Some reflections on the beginnings and development of statistics in "your father's NIH". Statistical Science 1997;12:82-7.
60. Dorn HF. Methods of analysis for follow-up studies. Hum Biol 1950;22:238-48.
61. Dorn HF. Methods of measuring incidence and prevalence of disease. Am J Public Health Nations Health 1951;41:271-8.
62. Aronowitz R. Les maladies ont-elles un sens? Paris: Synthélabo, 1999.
63. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J. Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. Ann Intern Med 1961;55:33-50.
64. Dawber TR. Summary of recent literature regarding cigarette smoking and coronary heart disease. Circulation 1960;22:164-6.
65. Doyle JT, Dawber TR, Kannel WB, Heslin AS, Kahn HA. Cigarette smoking and coronary heart disease. Combined experience of the Albany and Framingham studies. N Engl J Med 1962;26:796-801.
66. Stamler J. Established major coronary risk factors: historical review. In: Marmot M, Elliott P, eds. Coronary heart disease epidemiology, from aetiology to public health. Oxford: Oxford University Press, 1992.
67. Last JM. A dictionary of epidemiology. Oxford: Oxford University Press, 1995.
68. Dawber TR, Kannel WB. Application of epidemiology of coronary heart disease to medical practice. Mod Med 1962;30:85-101.
69. Kannel WB, Kagan A, Dawber TR, Revotskie N. Epidemiology of coronary heart disease. Implications for the practicing physician. Geriatrics 1962;17:675-90.
70. Cornfield J. Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. Fed Proc 1962;21(4)Pt 2:58-61.
71. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis 1967;20:511-24.
72. Cornfield J, Gordon T, Smith W. Quantal response curves for experimentally uncontrolled variables. Bull Int Stat Inst 1961;38:97-115.
73. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika 1967;54:167-79.
74. Gordon T, Sorlie P, Kannel WB. Coronary heart disease, atherothrombotic brain infarction, intermittent claudication – multivariate analysis of factors related to their incidence In: Kannel W, Gordon T, eds. The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease, Section 27. Washington: DHEW publication (NIH), 1971.
75. Coronary Risk Handbook: Estimating the Risk of Coronary Heart Disease in Daily Practice. Dallas: American Heart Association, 1973.
76. MacGee D, Gordon T. The results of the Framingham study : an epidemiological investigation of cardiovascular disease. In: Kannel W, Gordon T, eds. The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease, Section 31. Washington: DHEW publication (NIH), 1976.
77. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol 1976;38:46-51.